Applied and Environmental Microbiology, Aug. 2010, p. 5636–5638 Vol. 76, No. 16
0099-2240/10/$12.00 doi:10.1128/AEM.00303-10

# Integrated Microbial Survey Analysis of Prokaryotic Communities for the PhyloChip Microarray▿†

Michael C. Schatz,[1,2]‡ Adam M. Phillippy,[1,2]‡ Pawel Gajer,[1] Todd Z. DeSantis,[3] Gary L. Andersen,[3] and Jacques Ravel[1]*

*Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland 21201[1]; Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland 20742[2]; and Center for Environmental Biotechnology, Lawrence Berkeley National Laboratory, Berkeley, California 94720[3]*

Received 3 February 2010/Accepted 15 June 2010

**PhyloTrac is an integrated desktop application for analysis of PhyloChip microarray data. PhyloTrac combined with PhyloChip provides turnkey and comprehensive identification and analysis of bacterial and archaeal communities in complex environmental samples. PhyloTrac is free for noncommercial organizations and is available for all major operating systems at http://www.phylotrac.org/.**

The PhyloChip is a low-cost Affymetrix GeneChip microarray, developed at Lawrence Berkeley National Laboratory (LBNL), designed to detect and quantify abundance of bacterial and archaeal taxa using signature probes targeting all known 16S rRNA gene sequences. The second generation of the PhyloChip microarray targets nearly 9,000 operational taxonomic units (OTUs), with an average of 24 probes, each 25 bp long, and the upcoming third-generation PhyloChip application will target an even larger number of OTUs. Multiple, complex environments have been successfully analyzed using the PhyloChip microarray, including, among others, air (2), soil (1), the human lung (6), and the gut (9). PhyloChip microarrays are manufactured by Affymetrix, but to date, analysis has been available only from within LBNL, limiting the accessibility of the technology. PhyloTrac addresses this limitation by providing a standardized analysis package for the PhyloChip microarray, including microarray normalization, OTU quantification, multiple interactive visualizations, and integrated analytics.

**OTU identification and quantification.** PhyloTrac is entirely self-contained, requiring as input only the Affymetrix CEL files containing the raw PhyloChip probe intensities. All other required microarray design and taxonomy information is bundled into the application. From the CEL files, PhyloTrac performs normalization and scoring of the data using methods previously described by DeSantis et al. (4). This method has been successfully applied and validated by 16S rRNA gene clone library sequencing in prior studies utilizing the PhyloChip microarray (1, 2, 6).

Briefly, the PhyloChip data are first processed for background subtraction and normalization based on the background signal level and the observed intensity of synthetic probes for known spike-in concentrations. Widely variable or otherwise defective hybridizations are detected through quality control tests available within the application. Then, specific probe hybridization is estimated by comparing the intensity of a perfect-match (PM) probe to that of a paired mismatch (MM) probe containing a single difference at the central position. An OTU is considered present if the positive fraction (PF) of probes targeting that OTU is greater than the user-selected threshold (default, 95%). A hybridization score is computed for each positive OTU as the trimmed mean intensity of its PM probes. The hybridization score has been shown to have a strong linear correlation ($r = 0.917$) with the abundance of the DNA target present in the sample (5). After processing, PhyloTrac stores the results in an exchangeable binary format for further analysis within the application. Alternatively, OTU PF values and hybridization scores may be exported to a tab delimited file, such as one used for further analysis within R/Bioconductor (7), or in a format compatible with UniFrac (8) for additional statistical analyses of microbial community structure. 16S rRNA gene sequencing data can be codisplayed in all views to confirm or supplement the hybridization data.

**Visualization and analysis.** In addition to standardizing the scoring of PhyloChip microarray data, PhyloTrac provides powerful visualization and analysis features for exploring microbial diversity. Figure 1 recreates several views from the urban aerosol study by Brodie et al. (2), in which the microbial composition of the air in San Antonio and Austin, TX, was monitored for 17 weeks. Among other findings, this study discovered the regular presence of pathogenic OTUs, correlated aerosol microbial diversity with environmental conditions, and identified a set of 80 core subfamilies consistently detected in San Antonio. Using PhyloTrac's interactive and fully synchronized views of the data, including taxonomic trees, heat maps, hierarchical clustering, time series (parallel coordinates), difference plots, and multidimensional scaling scatter plots, it is possible to reproduce their analysis in just a few steps (see the supplemental material).

The primary PhyloTrac window displays a taxonomic tree of the OTUs detected by the PhyloChip microarray, with mean

* Corresponding author. Mailing address: Institute for Genome Sciences, University of Maryland School of Medicine, BioPark II, Room 611, 801 West Baltimore Street, Baltimore, MD 21201. Phone: (410) 706-5674. Fax: (410) 706-1482. E-mail: jravel@som.umaryland.edu.

† Supplemental material for this article may be found at http://aem.asm.org/.

‡ M.C.S. and A.M.P. contributed equally to this work.

▿ Published ahead of print on 25 June 2010.

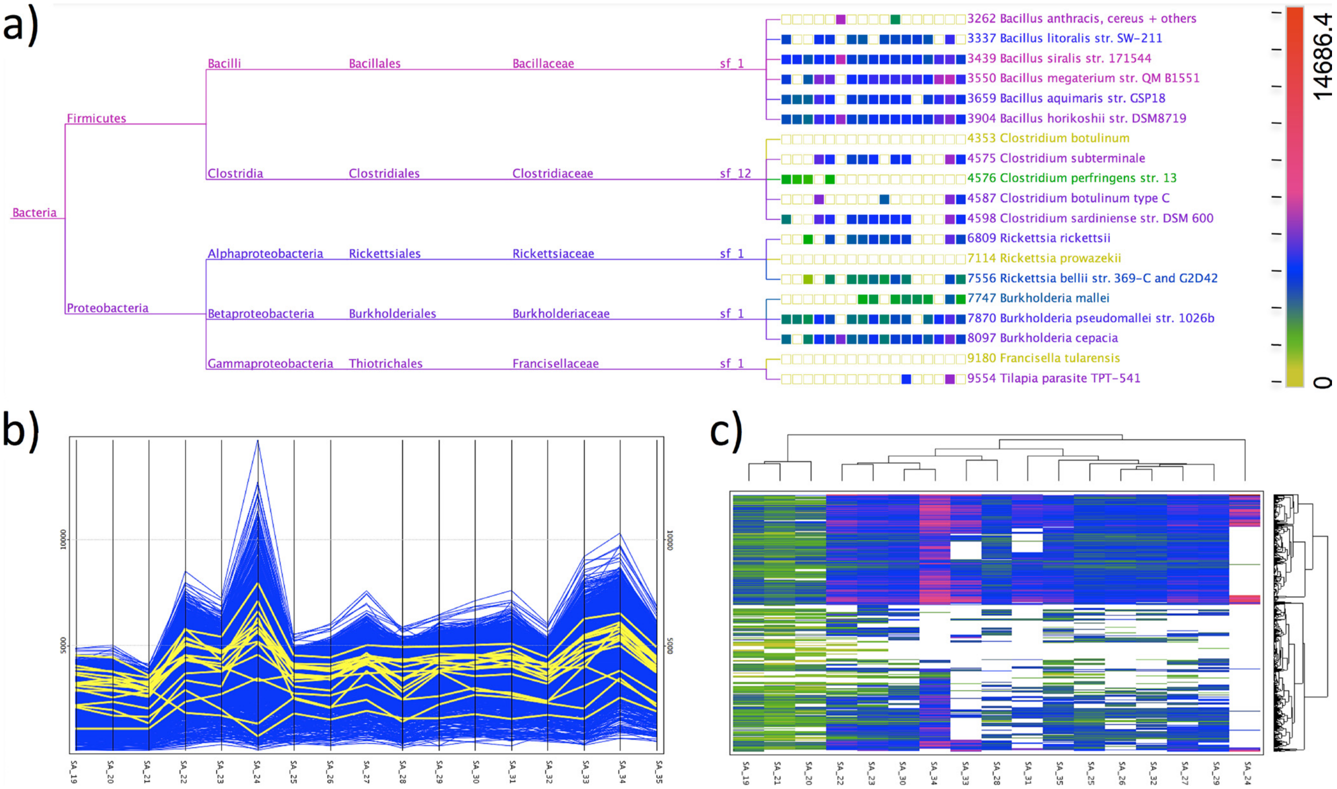
a)
Bacteria
Firmicutes
Proteobacteria
Bacilli
Bacillales
Bacillaceae
sf_1
Clostridia
Clostridiales
Clostridiaceae
sf_12
Alphaproteobacteria
Rickettsiales
Rickettsiaceae
sf_1
Betaproteobacteria
Burkholderiales
Burkholderiaceae
sf_1
Gammaproteobacteria
Thiotrichales
Francisellaceae
sf_1
3262 Bacillus anthracis, cereus + others
3337 Bacillus litoralis str. SW-211
3439 Bacillus siralis str. 171544
3550 Bacillus megaterium str. QM B1551
3659 Bacillus aquimaris str. GSP18
3904 Bacillus horikoshii str. DSM8719
4353 Clostridium botulinum
4575 Clostridium subterminale
4576 Clostridium perfringens str. 13
4587 Clostridium botulinum type C
4598 Clostridium sardiniense str. DSM 600
6809 Rickettsia rickettsii
7114 Rickettsia prowazekii
7556 Rickettsia bellii str. 369-C and G2D42
7747 Burkholderia mallei
7870 Burkholderia pseudomallei str. 1026b
8097 Burkholderia cepacia
9180 Francisella tularensis
9554 Tilapia parasite TPT-541
14686.4
0
b)
10000
5000
SA_19
SA_20
SA_21
SA_22
SA_23
SA_24
SA_25
SA_26
SA_27
SA_28
SA_29
SA_30
SA_31
SA_32
SA_33
SA_34
SA_35
c)
SA_19
SA_21
SA_20
SA_22
SA_23
SA_30
SA_34
SA_33
SA_28
SA_31
SA_35
SA_25
SA_26
SA_32
SA_27
SA_29
SA_24

FIG. 1. (a) Taxonomic tree window showing the hybridization intensity of selected OTUs for each sample, ordered vertically by taxonomy and horizontally by sample. Labels are shown at each level of the taxonomy, from domain (left) to OTU (right). (b) Time series window showing the change in hybridization intensity for each OTU across all samples, with the selected OTUs shown in yellow. (c) Heat map and clustering window showing a hierarchical clustering of both the OTUs (rows) and the samples (columns).

intensities for each detected OTU displayed as a heat map of samples at the leaves of the tree. The user may dynamically filter the tree to hide low-abundance or borderline OTUs below user-specified intensity or PF thresholds, search and filter by keyword or clade, or summarize the analysis at any level of the tree from phylum to species. In Fig. 1a, the tree has been filtered to display the abundance of 19 select pathogenic or near-neighbor OTUs in San Antonio throughout the 17 weeks. In accordance with reference 2, the regular presence of phylogenetic near-neighbors to *Bacillus anthracis*, *Burkholderia pseudomallei*, and *Clostridium botulinum* is evident. In a second synchronized window, these OTUs are displayed in a time series plot, displaying their change in abundance over time, highlighted in yellow, in relation to all other OTUs, colored blue, and showing a spike in abundance in week 24 and a depression in weeks 19 to 21 (Fig. 1b). In a third heat map window, users can hierarchically cluster both OTUs and samples using any of the standard distance and linkage methods from the integrated C clustering library (3) (Fig. 1c). Clustering the OTUs reveals the organisms with similar patterns of abundance across samples, while clustering the samples can confirm sample replicates or reveal biologically similar environments. For example, in Fig. 1c, the top split in the sample dendrogram clusters samples from weeks 19 to 21, which Brodie et al. correlate with significant differences in environmental conditions during those 3 weeks. Another integrated visualization displays interactive bar plots of differential OTU intensities, where the heights of the bars display either the absolute or the relative difference in intensity for each OTU between a pair of samples (not shown). As intensity has been strongly correlated with abundance, a difference plot can be an effective summarization of relative changes in abundance between environments.

Synchronized selection and filtering afford users the unique ability to seamlessly navigate between multiple views of the data. For example, users can select a cluster in the hierarchical clustering window and simultaneously view the selected organisms in the taxonomy tree or time series display, immediately relating their phylogenetic, environmental, and temporal relationships. This deep level of integration makes it possible to quickly and deeply analyze complex microbial environments.

PhyloTrac is implemented in C++, using the cross-platform Qt development framework (Nokia Corporation), and runs natively on Macintosh, Linux, and Windows operating systems.

Installing the R statistical package (http://www.r-project.org/) unlocks multidimensional scaling functionality and an integrated R console for additional analysis. PhyloTrac binaries and user manual are available at the PhyloTrac home page (http://www.phylotrac.org/).

This work was supported by the University of Maryland School of Medicine.

## REFERENCES

1. **Brodie, E. L., T. Z. Desantis, D. C. Joyner, S. M. Baek, J. T. Larsen, G. L. Andersen, T. C. Hazen, P. M. Richardson, D. J. Herman, T. K. Tokunaga, J. M. Wan, and M. K. Firestone.** 2006. Application of a high-density oligonucleotide microarray approach to study bacterial population dynamics during uranium reduction and reoxidation. Appl. Environ. Microbiol. **72:**6288–6298.
2. **Brodie, E. L., T. Z. DeSantis, J. P. Parker, I. X. Zubietta, Y. M. Piceno, and G. L. Andersen.** 2007. Urban aerosols harbor diverse and dynamic bacterial populations. Proc. Natl. Acad. Sci. U. S. A. **104:**299–304.
3. **de Hoon, M. J., S. Imoto, J. Nolan, and S. Miyano.** 2004. Open source clustering software. Bioinformatics **20:**1453–1454.
4. **DeSantis, T. Z., E. L. Brodie, J. P. Moberg, I. X. Zubieta, Y. M. Piceno, and G. L. Andersen.** 2007. High-density universal 16S rRNA microarray analysis reveals broader diversity than typical clone library when sampling the environment. Microb. Ecol. **53:**371–383.
5. **DeSantis, T. Z., C. E. Stone, S. R. Murray, J. P. Moberg, and G. L. Andersen.** 2005. Rapid quantification and taxonomic classification of environmental DNA from both prokaryotic and eukaryotic origins using a microarray. FEMS Microbiol. Lett. **245:**271–278.
6. **Flanagan, J. L., E. L. Brodie, L. Weng, S. V. Lynch, O. Garcia, R. Brown, P. Hugenholtz, T. Z. DeSantis, G. L. Andersen, J. P. Wiener-Kronish, and J. Bristow.** 2007. Loss of bacterial diversity during antibiotic treatment of intubated patients colonized with Pseudomonas aeruginosa. J. Clin. Microbiol. **45:**1954–1962.
7. **Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang.** 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. **5:**R80.
8. **Hamady, M., C. Lozupone, and R. Knight.** 2010. Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. ISME J. **4:**17–27.
9. **Ivanov, I. I., K. Atarashi, N. Manel, E. L. Brodie, T. Shima, U. Karaoz, D. Wei, K. C. Goldfarb, C. A. Santee, S. V. Lynch, T. Tanoue, A. Imaoka, K. Itoh, K. Takeda, Y. Umesaki, K. Honda, and D. R. Littman.** 2009. Induction of intestinal Th17 cells by segmented filamentous bacteria. Cell **139:** 485–498.